Multiagent Data Collection in Lycos

Richard Green and Sangam Pant

We proceed from a relatively weak notion of agency in cooperating software system components to describe the data collection activities of the Lycos Internet information retrieval services. The characteristics of agency of interest include autonomy, social ability, reactivity, and proactivity.

The original and simple Lycos spiders, once implemented in Perl, have evolved into a true multiagent system of cooperating components that can visit and analyze more than 10,000,000 Web pages each day. There are three kinds of cooperating components:

Spiders are independent software agents that crawl the Web to gather information. In the Lycos data collection system, the independent multiple processes are also multithreaded. The spiders individual spider threads communicate directly with another component called the Update Server (US). They get their marching orders—the set of URLs they are to visit—from the US and they pass back any discovered hyperlinks that the US can then parcel out to be visited in turn.

The **URL server** manages which servers and pages are to be visited by the spiders. Its job is to give each spider a list of URLs to visit and to receive from each spider the data it collects about links the spider may discover as it travels the Web. It provides a mechanism for controlling the rate at which spiders work and the environs in which they operate, providing command and control for the spiders. One Update Server manages the working of numerous spiders.

The **Catalog Update Server** (CUS) receives data from spiders, prepares it for indexing and stores it in a repository.

The three independent components of this system were developed as individual programs and communicate as needed to gather and analyze specific kinds of Web-based hyperlinked documents. The Lycos data collection system is clearly a distributed computing system residing and operating within an internetworked environment. It can also be viewed as a multiagent system because its components act as independent agents interoperating with one another to achieve greater reach.

There are advantages to building the Lycos data collection system as a multiagent system. We could have chosen to build all of the capabilities of the spider US—CUS complex into the spider. Indeed, the original Lycos spider did perform all of these functions itself. There are several reasons why choosing a multiagent design made sense for Lycos.

- *Better management of site visitations.* Web robots have been known to tie up the resources of smaller Web servers by making too frequent visits as they discover links. Making the generation of URL lists for spiders to visit the responsibility of a second cooperating agent allows us to cover large amounts of the Web without unduly taxing other Web server resources.
- *Ability to scale with the Web.* Merely duplicating many spiders is very resource intensive. Creating multithreaded spiders of very small process footprint that communicate with the other components of the multiagent system allows economical coverage of very large parts of the Web and permits us to keep pace with growth as more HTTP servers come online.
- *Batching of computationally expensive processing.* Moving some of the document parsing and data transformations into the CUS permits us to devote specific hardware and storage resources to managing very large amounts of data. Spiders can gather data very quickly; certain operations like language and topic classification are computationally intense and better viewed as batch process. CUS can save up data from many spiders and process it in an efficient manner to prepare it for indexing.

The Lycos data collection system is truly a good example of multiagent systems in large-scale information environments. C

SANGAM PANT (spant@lycos.com) is the vice president of networking and site integration at Lycos, Inc., Waltham, Mass. He is the contact author for this article.

The original Lycos spiders have evolved into a multiagent system of cooperating components that can visit and analyze more than 10,000,000 Web pages each day.